Large Language Models Can Be Easily Distracted by Irrelevant Context

Year: 2023 | *Citations: 777* | *Authors: Freda Shi, Xinyun Chen, Kanishka Misra*

Abstract

Large language models have achieved impressive performance on various natural language processing tasks. However, so far they have been evaluated primarily on benchmarks where all information in the input context is relevant for solving the task. In this work, we investigate the distractibility of large language models, i.e., how the model problem-solving accuracy can be influenced by irrelevant context. In particular, we introduce Grade-School Math with Irrelevant Context (GSM-IC), an arithmetic reasoning dataset with irrelevant information in the problem description. We use this benchmark to measure the distractibility of cutting-edge prompting techniques for large language models, and find that the model performance is dramatically decreased when irrelevant information is included. We also identify several approaches for mitigating this deficiency, such as decoding with self-consistency and adding to the prompt an instruction that tells the language model to ignore the irrelevant information.